Measurement and Analysis of the Swarm Social Network With Tens of Millions of Nodes

Chen, Yang

2018-01-23

Chen , Y , Hu , J , Zhao , H , Xiao , Y & Hui , P 2018 , ' Measurement and Analysis of the Swarm Social Network With Tens of Millions of Nodes ' , IEEE Access , vol. 6 , pp. 4547-4559 . https://doi.org/10.1109/ACCESS.2018.2789915

http://hdl.handle.net/10138/237065
https://doi.org/10.1109/ACCESS.2018.2789915

unspecified
publishedVersion

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

received October 6, 2017, accepted November 28, 2017, date of publication January 23, 2018,
date of current version February 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2789915

Measurement and Analysis of the Swarm Social Network With Tens of Millions of Nodes

YANG CHEN[1,2,3], (Senior Member, IEEE), JIYAO HU[1,2,3], HAO ZHAO[1,2,3], YU XIAO[4], (Member, IEEE), and PAN HUI[5,6], (Fellow, IEEE)

[1]School of Computer Science, Fudan University, Shanghai 200433, China
[2]Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, Shanghai 200433, China
[3]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710126, China
[4]Department of Communications and Networking, Aalto University, 02150 Espoo, Finland
[5]Department of Computer Science, University of Helsinki, 00100 Helsinki, Finland
[6]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong

Corresponding author: Yang Chen (chenyang@fudan.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61602122 and Grant 71731004, in part by the Natural Science Foundation of Shanghai under Grant 16ZR1402200, in part by the Shanghai Pujiang Program under Grant 16PJ1400700, in part by the Academy of Finland under Grant 268096, in part by the General Research Fund from the Research Grants Council of Hong Kong under Grant 26211515 and Grant 16214817.

ABSTRACT Social graphs have been widely used for representing the relationship among users in online social networks (OSNs). As crawling an entire OSN is resource- and time-consuming, most of the existing works only pick a sampled subgraph for study. However, this may introduce serious inaccuracy into the analytic results, not to mention that some important metrics cannot even be calculated. In this paper, we crawl the entire social network of Swarm, a leading mobile social app with more than 60 million users, using a distributed approach. Based on the crawled massive user data, we conduct a data-driven study to get a comprehensive picture of the whole Swarm social network. This paper provides a deep analysis of social interactions between Swarm users, and reveals the relationship between social connectivity and check-in activities.

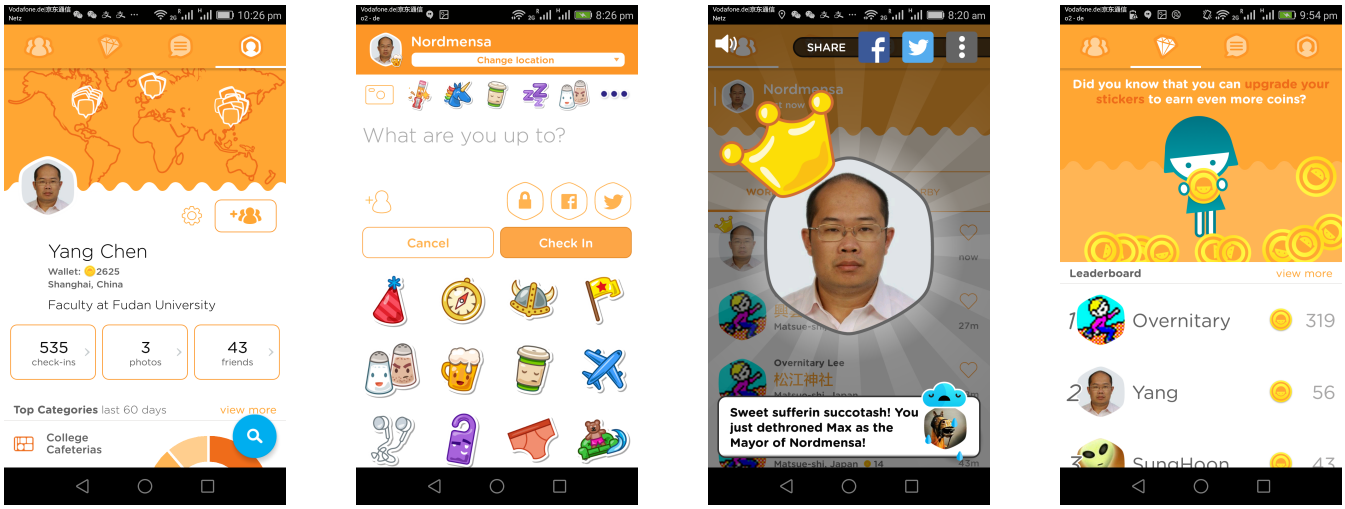
INDEX TERMS Social network analytics, Swarm app, social graph, user-generated contents.

I. INTRODUCTION

Nowadays online social networks (OSNs) have become extremely popular around the world, and have attracted billions of users [23]. To gain a deep understanding of an OSN, social graphs have been widely used for describing and analyzing the interactions between users [12], [13], [21], [26], [32], [34], [47], [50], [54]. They have been adopted in studies of data placement [22], [56], information diffusion [27], cloud computing [5], trustworthy distributed computing [33], and social data delivery [51].

Most of the mainstream OSNs have applied a per-IP address rate limit, which controls for example how many requests are allowed to be sent per hour from a unique IP address. Therefore, it would take lots of time and network resources to crawl the entire social graph of a large-scale OSN. Due to this, most of the existing works have chosen to pick a sampled subgraph for study, unless the authors have direct access to the back-end of the OSN service [1], [34], [47], [54]. Examples of sampling algorithms include Breadth-First Search (BFS) [48] and Metropolis-Hastings Random Walk (MHRW) [13]. Because a sampled subgraph can only provide a partial view of an OSN, the analysis results based on the subgraphs may be biased and may not precisely represent all the key features of the OSN [48].

This work aims at providing a comprehensive view of a mainstream OSN that consists of tens of millions nodes. We choose Foursquare's Swarm app, a dedicated mobile social app with more than 60 million users around the world. In order to speed up the data collection, we launched a number of servers with different IP addresses. These servers collaborated with each other to fetch the user data, using the crowd crawling framework [10]. We were able to crawl the whole Swarm network within 40 days in 2015. For each user, our data set records her profile page and a complete list of her friends. Based on the data set, we create the entire social graph of Swarm, and calculate the key graph metrics, including the degree, clustering coefficient, assortativity, PageRank, connected components and communities. In addition, we study the app usage, taking the check-in function as an example, and propose a classification algorithm



(a) (b) (c) (d)

FIGURE 1. Screenshots of the Swarm App. (a) Profile. (b) Check-in. (c) Mayorship. (d) Virtual coins.

for predicting active users. We have made the anonymized data set publicly available via `https://github.com/chenyang03/Swarm_dataset`.

Our contributions are summarized below.

- We perform a demographic analysis that investigates the composition of Swarm users from different aspects, including gender, location and their privacy concerns. This gives an informative overview of Swarm users.
- We create the entire social graph of Swarm without directly accessing the back-end of the Swarm site, and conduct a comprehensive analysis of the social interaction between Swarm users. Compared with the analysis results based on subgraphs, our work demonstrates the need for studying the entire social graph.
- Our work discovers the predictability of user behavior by investigating Swarm users' check-in activities. We have found that a user's social connections and profile can be used to accurately predict a user's activeness in terms of the number of check-ins.

The rest of this paper is organized as follows. We discuss the background of the Swarm app, and the procedure of data collection in Section II. We analyze the collected massive Swarm data in Section III. We review the related work in Section IV, before we conclude the work in Section V.

II. BACKGROUND AND DATA COLLECTION

Since 2009, Foursquare [38], [39], [44] has been a leading site for the combination of location-based services (LBS) and mobile social networking. Different from traditional OSNs, such as Facebook and LinkedIn, all activities on Foursquare are related to certain venues, a.k.a., points of interest. In May 2014, the original Foursquare was split into two apps, i.e., Foursquare City Guide and Swarm. The brand new Foursquare City Guide app acts as a Yelp-like platform for discovering and posting comments on new places. Differently, most of the classic functions of the original Foursquare app, such as check-in and mayorship, are integrated into the new Swarm app. Swarm focuses on serving mobile users by engaging them in location-centric activities. Swarm users can share their locations through the social network. Note that Swarm does not offer any interface for desktop users to conduct check-ins or make friends. Fig. 1(a) shows the profile page of a Swarm user. If you click the "friends" box, you can obtain the full friend list of this user.

The primary way of content publishing in Swarm is known as "check-ins". As shown in Fig. 1(b), a user can conduct a check-in to publish the real-time location, and this information will be shown on the timelines of his Swarm friends. Accordingly, a Swarm user can browse the location history of each of his friends. To make a check-in informative, a user can attach a "sticker" to the check-in to express his feeling or indicate what he is doing, for example, drinking a beer. By using the cross-site linking function [8], the check-ins can be automatically shared to other social networks including Facebook and Twitter.

There are two types of incentives in Swarm to encourage users to check in more often. A user can play location-based games with other users, e.g., competing for the "mayorship" crown of a certain place by visiting the place more frequently. In Fig. 1(c), we can see a user has seized the mayorship crown after conducting a check-in. Meanwhile, Swarm provides virtual coins for users according to their activities. The more active the user is, the more coins she would earn. Fig. 1(d) shows a "leaderboard" which ranks a user and his friends based on the amount of coins. Since June 2016, virtual coins can be used for getting discounts when conducting check-ins at certain businesses.[1] This provides

[1]https://www.theverge.com/2016/9/13/12896836/foursquare-swarm-deals-update-version

a viable economic incentive for Swarm users to undertake more check-ins.

We use the social graph to analyze the social interactions between massive Swarm users. To obtain a snapshot of the entire social graph of Swarm, we need to crawl the data of all 60+ million Swarm users. This task is not trivial, because Swarm applies a strict rate limit for each IP address, which prevents us from obtaining massive user data in a short time. For each user's data, we use the official Swarm API to conduct the crawling. Each Swarm user has a unique numeric ID, and the IDs are assigned successively. Therefore, we can register a new Swarm account to get the maximum ID, denoted by max_uid. In order to deal with the IP-based rate limit, we launched 40 crawlers in parallel. Each crawler was deployed on a virtual instance of the Microsoft Azure platform, with a unique IP address. We split the whole ID range $[1, max\_uid]$ evenly into 40 chunks, with each crawler taking care of one chunk.

From August 1 to September 10 in 2015, we crawled all 62.6 million Swarm users' profile pages and friend lists. Note that we respect the privacy of Swarm users that we only crawled the publicly-accessible information. Based on the crawled friend lists of all Swarm users, we model the entire Swarm network using an undirected social graph $G = (V, E)$. V is the set of all Swarm users, and E is the set of social connections between users. A node in V represents a user, and an edge in E represents a social connection. For any two nodes $v_1 \in V$ and $v_2 \in V$, an edge $e \in E$ between them indicates that these two users are friends in the Swarm network. The degree of a node in G indicates the number of friends the corresponding user has. The clustering coefficient (CC) of a node in G denotes the fraction of pairs of the node's neighbors that are directly connected to each other. The clustering coefficient of node i is defined as follows:

$$C_i = \begin{cases} \dfrac{2e_i}{k_i(k_i-1)} & k_i > 1 \\ 0, & k_i \in \{0, 1\} \end{cases} \quad (1)$$

where k_i is the degree of node i, and e_i is the number of connected pairs between any two neighbors of i.

The social graph G we create has 62,602,899 nodes and 777,559,146 edges. To the best of our knowledge, this is the first measurement-based work that analyzes the entire social graph of a mainstream mobile social network without directly accessing the back-end of the site. Our method can be adopted by scholars and third-party application providers to crawl and analyze the entire social graph of Swarm or other similar social networks.

III. DATA ANALYSIS

In this section, we study the behavior of Swarm users using the data set described in Section II. Based on the user profiles, we first conduct a demographic analysis in Section III-A, and then demonstrate the necessity of getting the entire social graph in Section III-B. After that, we measure a series of key graph metrics in Section III-C, and investigate the predictability of the check-in behavior in Section III-D.

A. DEMOGRAPHIC ANALYSIS

To understand the composition of Swarm users, we analyze several key information fields of user profiles. We first look at the gender distribution. Among all Swarm users, 51.07% of them are male, 41.43% are female, and 7.50% do not want to disclose their gender information.

Regarding user location, 89.69% of Swarm users have filled out the optional "location" field. We use the Google Geocoding API[2] to interpret the country information from the user input. According to our study, 24.13%, 11.68%, 7.17% and 5.90% of Swarm users come from the USA, Turkey, Indonesia and Brazil, respectively. The users from these four countries cover about half of the entire Swarm population.

A Swarm user is allowed to link her Swarm account with her Facebook and/or Twitter accounts. By enabling this cross-site linking function [8], a user can publish her check-ins on other OSNs automatically. Meanwhile, she can import the friend lists from the linked OSN accounts [55]. We can see that 56.76% of users have linked their Swarm profiles to their Facebook and/or Twitter accounts.

We also group the users according to their privacy concerns. In addition to the mandatory information fields, there are five optional fields in a Swarm user's profile, i.e., last name, gender, profile photo, home location, and biography. We consider a user as an "open" user, if she has filled out at least four out of these five fields. Similarly, we regard a user as "cautious", if at least 4 fields are left empty. Accordingly, 1.70% of users are "open", 40.01% are "cautious", and the others are grouped into "other".

B. SIGNIFICANCE OF GETTING THE ENTIRE SOCIAL GRAPH

In this subsection, we explain why getting the entire social graph is essential for conducting an unbiased analysis of the social network. In the literature, people usually use a sampled subgraph, for example, a small portion of nodes and the corresponding edges. The well-known breadth-first search (BFS) algorithm starts from adding a selected user to a queue. In each step, we pick the first user from the head of the queue, download her profile, and obtain a list of her friends. These friends, if they have not been crawled, will be added to the queue. This procedure will be repeated until the queue is empty, or until we have collected enough user data. BFS has been widely used, such as in [14], [15], and [50], given its simplicity. However, such a sampling method is biased, and has a higher probability of including more nodes with relatively high degrees in the data set. As a result, the sampled users cannot represent the entire user population. On the other hand, scholars have proposed some "unbiased sampling" algorithms, such as Metropolis-Hastings Random

[2]https://developers.google.com/maps/documentation/geocoding

TABLE 1. Mean and variance of the average degrees and clustering coefficients of different subgraphs.

Metric		BFS (1%)	BFS (5%)	BFS (10%)	MHRW (1%)	MHRW (5%)	MHRW (10%)	LCC	Entire
Degree	Mean (μ)	197.60	161.51	135.01	44.87	44.86	44.88	43.56	24.84
	Variance (σ^2)	4208.66	1412.26	785.20	2.94	0.51	0.29	N/A	N/A
Clustering Coefficient	Mean (μ)	0.08	0.09	0.09	0.14	0.14	0.14	0.14	0.08
	Variance (σ^2)	3.11×10^{-4}	2.21×10^{-4}	1.61×10^{-4}	9.16×10^{-6}	1.81×10^{-6}	1.15×10^{-6}	N/A	N/A

Walk (MHRW) [13]. MHRW is a Markov-Chain Monte Carlo (MCMC) algorithm that gets a random sample of nodes according to the degree distribution of the nodes. Unfortunately, MHRW only works on connected graphs.

We execute the BFS and MHRW algorithms on G, respectively, to sample different subgraphs. To validate whether a sampled subgraph can represent the entire network, we use two metrics, degree and clustering coefficient, to compare the analysis results between the entire G and the subgraphs obtained by BFS and MHRW, respectively. We start BFS and MHRW from a randomly selected node within the largest connected component (LCC). Since the Swarm network is not a connected graph, the sampling results will not include any nodes from other connected components. For both BFS and MHRW, we examine the cases where 1%, 5% and 10% of all nodes are sampled, respectively. For each algorithm with a specified sampling percentage, we run it 500 times independently and report the mean and the variance of both the average degree and clustering coefficient metrics in Table 1.

We find that neither BFS nor MHRW could accurately represent the average degree of the entire Swarm network. In the case of BFS, the average degree of the sampled nodes is much larger than that of the LCC, not to mention the entire Swarm network. The value of average degree gets higher, if less samples are collected. For MHRW, we find that the mean values of average degrees of sampled subgraphs remain stable, regardless of the size of the sample data, and is very close to the average node degree of the LCC. These results confirm that MHRW can accurately capture the average degree of a connected component. However, this value is still nearly two times of the entire Swarm graph. A similar phenomenon can be observed for the measurements of the mean values of average clustering coefficients. MHRW can characterize the average clustering coefficient of the LCC precisely, but not the entire Swarm network.

Regarding the variance, the MHRW algorithm can always achieve a small variance for both average degrees and clustering coefficients. In other words, MHRW can obtain the average degree and clustering coefficient of the LCC in an accurate and stable way. In contrast, the BFS algorithm results in a much larger variance.

In addition to the inaccuracy caused by the node sampling, some widely used methods in graph analysis, such as community detection [2] and PageRank calculation [40], require the information of all nodes and edges of the graph. Therefore, they cannot be undertaken, given a sampled subset of the graph. Therefore, our efforts on getting the entire social graph are necessary for obtaining a comprehensive view of the Swarm network. In Section III-C, to understand the social interactions between users, we will study the entire Swarm social graph from different angles.

C. ANALYZING THE ENTIRE SWARM GRAPH

To analyze the entire Swarm graph with tens of millions of nodes, we use the Stanford Network Analysis Platform (SNAP) [29], which is a general purpose library for social network analysis. It is implemented in C++ and can scale up to large networks with millions of nodes. We are interested in the following graph metrics, i.e., degree, clustering coefficient, assortativity, PageRank and connected components. Also, we analyze the communities formed by users.

1) DEGREE AND CLUSTERING COEFFICIENT

Fig. 2(a) shows the probability density function (PDF) of the degrees of all nodes. We can see that 37.69% of nodes have a degree larger than 5, while 30.19% of nodes have a degree larger than 10. Although the average node degree in Swarm is 24.84, there are still 26,607,755 nodes (42.5% of all nodes) which do not have any friend.

Earlier results have shown that in most of the representative OSNs, such as Facebook [50], Orkut [32], and Flickr [32], the degree distribution can be approximated by a power-law distribution. This finding has been widely used for studying natural and man-made phenomena. To figure out the best fitting for the degree distribution of Swarm social graph, we compare the results of four well-known distributions, i.e., power law ($P(k) \propto Ck^{-\alpha}$) [9], power law with exponential cutoff ($P(k) \propto Ck^{-\alpha}e^{-\lambda k}$) [9], lognormal ($P(k) \propto e^{-\frac{(\ln k-\mu)^2}{2\tau^2}}$), and two-term exponential ($P(k) \propto ae^{bk} + ce^{dk}$). To compute the fitting parameters and the accuracy, we use the *cftool* (Curve Fitting Tool) in MATLAB 2016a, and quantify the fitting accuracy using a metric called *the coefficient of determination* (the R^2 value). The R^2 value is between 0 and 1. When the value is 1, it means the model fits the data perfectly. Based on our investigation, the degree distribution of Swarm can be approximated by a power-law model ($C = 13.65, \alpha = 0.8958$), with the corresponding R^2 value of 0.9809. According to Fig. 2(a), the statistical model fits the distribution very well.

We compare the average degree between different groups of users. According to Fig. 2(b), the average degree of male users is 27.28, while that of female users is 24.84. Male Swarm users tend to make more friends online than female ones. For the users who choose not to disclose the gender information, the average degree is only 9.13. They tend to

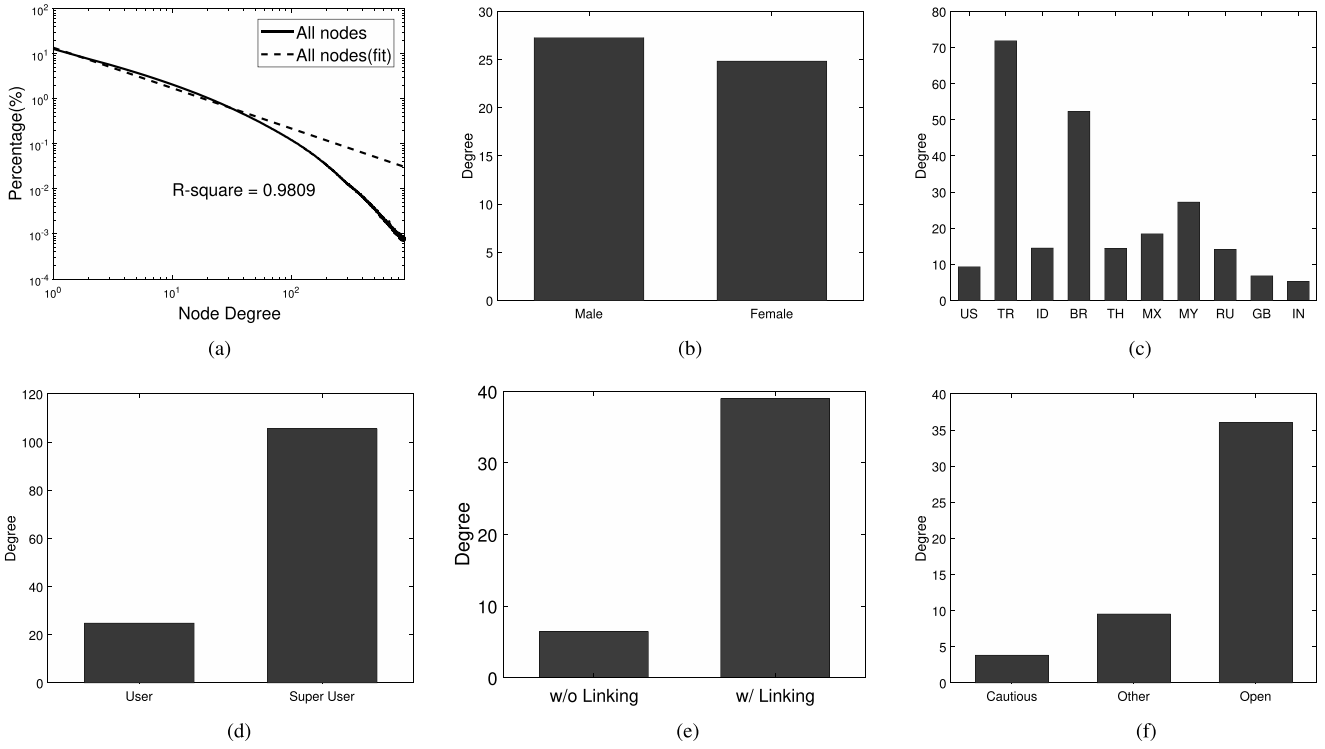


FIGURE 2. Degree. (a) Probability density function. (b) Gender. (c) Country. (d) User type. (e) Cross-site linking. (f) Privacy.

make fewer friends as they care more about privacy. We also group the users by home country. In Fig. 2(c), we show the average degrees of users from the top 10 countries with the highest Swarm population. On average, a user in Turkey has more than 70 friends. However, in United States, United Kingdom, and India, a user has less than 10 friends on average. Some most dedicated and passionate Swarm users are known as superusers.[3] These superusers are responsible for verifying the correctness of the location information. According to Fig. 2(d), being a superuser is an indicator of having a higher node degree. In addition, for a user who links her profile to Facebook/Twitter accounts (Fig. 2(e)), she has a higher chance to have more friends. For a user who cares more about her privacy, by contrast, she tends to have fewer friends in Swarm (Fig. 2(f)).

Fig. 3(a) shows the cumulative distribution function (CDF) of CC in the Swarm network. The average CC is 0.080. We also perform a group-based study on the average CC. According to Fig. 3(b), the average CC of male users is 0.080, while that of female users is 0.087. Therefore, male users have a slightly smaller average CC than female users. We also group the users by home country in Fig. 3(c). For Russian users, the average CC of a node is 0.145. However, for Indian users, the average CC of a node is 0.068. We also see that being a superuser (Fig. 3(d)) or linking the profile to Facebook/Twitter accounts (Fig. 3(e)) is an indicator of having a higher CC. Meanwhile, for a Swarm user who cares more about her privacy, she tends to have a smaller value of CC (Fig. 3(f)).

[3]Superusers in Foursquare/Swarm: https://support.foursquare.com/hc/en-us/articles/201066260-Superusers-SUs-

TABLE 2. Swarm v.s. other mainstream OSNs.

Network	Avg. Degree	Avg. CC
Swarm	24.84	0.08
Renren [54]	20.95	0.14
Cyworld [1]	31.64	0.16

There are several existing studies about social graphs of OSNs. Unfortunately, many of them are based on subgraphs, including [14], [15], [21], [32], and [50]. Because we need data sets that cover the entire networks for comparison, we select the following two mainstream OSNs, i.e., Cyworld and Renren. Ahn *et al.* [1] have analyzed the Cyworld data set provided by SK Communications, the provider of the Cyworld service. It is an anonymized snapshot of the entire Cyworld network captured in Nov. 2005. The data set contains 191 million social connections among 12 million users. Zhao *et al.* [54] have explored a data set provided by the Renren network. The data set covers the timestamped creation activities of all 19,413,375 users and 199,563,976 social connections of the Renren network during November 21, 2005 and December 31, 2007.

As shown in Table 2, our comparison focuses on two metrics, i.e., degree and clustering coefficient. The average degree of Swarm is much smaller than that of Cyworld, but larger than that of Renren. Regarding the average cluster-

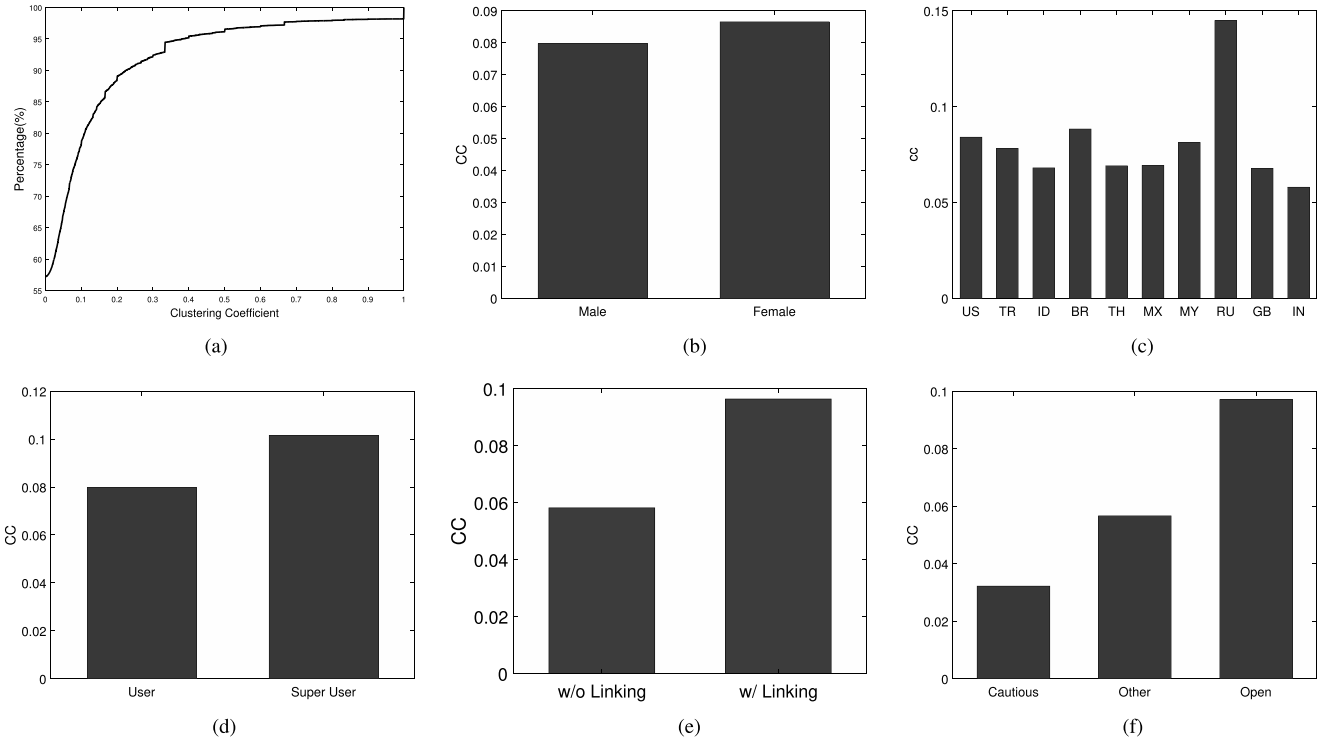


FIGURE 3. Clustering Coefficient. (a) Clustering coefficient distribution (CDF). (b) Gender. (c) Country. (d) User type. (e) Cross-site linking. (f) Privacy.

ing coefficient, the corresponding value of Swarm is much smaller than both Cyworld and Renren. Therefore, nodes in Swarm are not tightly connected with each other. We believe that it is because Swarm is not only for interacting with old friends, but also for playing location-centric games with strangers, and possibly making new friends.

2) ASSORTATIVITY

Degree assortativity r_{deg} [35] is a widely-used metric for quantifying the probability of connecting a node with other nodes with similar degrees. It is defined as the coefficient of the Pearson correlation between the degrees of any two connected nodes. The value of r_{deg} is between -1 and 1. A positive value indicates that there is a correlation between nodes of similar degrees. Such a graph is considered to follow an assortative mixing pattern. On the contrary, a graph with a negative r_{deg} is considered to have a disassortative mixing pattern. According to [36] and [37], most social networks show an assortative mixing pattern. The r_{deg} value of the Swarm network is 0.40, which is larger than the social networks studied in the existing literature including [1], [21], [32], [36], and [50]. Therefore, the Swarm social graph demonstrates a clear assortative mixing pattern of degree.

By grouping Swarm users based on demographic information such as gender and home country, we study the assortative mixing according to discrete characteristics. We first classify all users into m groups. Accordingly, we get an $m \times m$ symmetric mixing matrix E. As in [36], for such an undirected graph, each edge has a pair of unique X-end and Y-end. We define e_{ij} as the fraction that a randomly chosen edge is connected to a node of group i at its X-end and group j at its Y-end. We count the edges that are connected to group i at its X-end, and the ones that are connected to group j at its Y-end. The sums are denoted by $a_i = \sum_j e_{ij}$ and $b_j = \sum_i e_{ij}$, respectively. As in [35], the assortative coefficient can be calculated as

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} \tag{2}$$

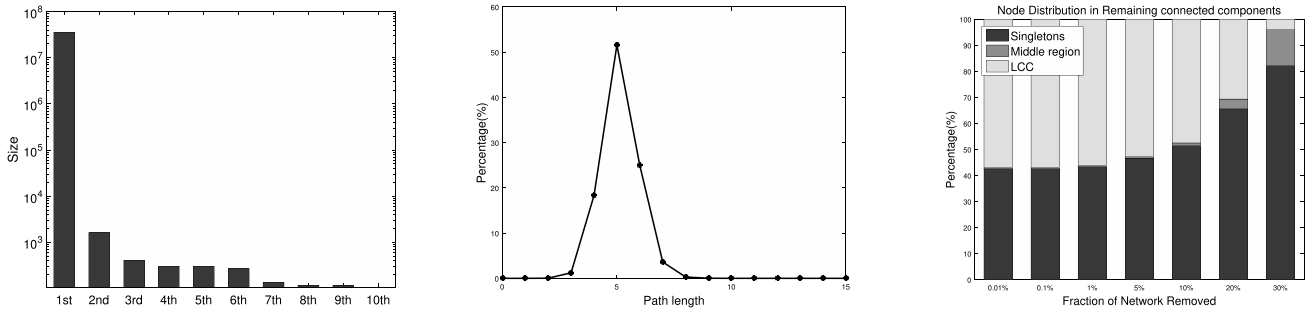
This metric is similar to but distinct from r_{deg}. Table 3 and Table 4 list the mixing matrix based on users' gender and home country information, respectively. For the gender-based mixing matrix, we ignore the users who have not provided their gender information. We find that a female user has a higher chance to make friend with a male user rather than another female user. Therefore, r_{gender} is as small as 0.104. For the home country-based mixing matrix, we only consider users coming from the top 10 countries with highest Swarm population. We observe that users tend to connect with other users coming from the same home country, as $r_{country}$ is as large as 0.865.

TABLE 3. Mixing matrix (gender).

	Male	Female
Male	0.496	0.192
Female	0.192	0.120

TABLE 4. Mixing matrix (country).

	US	TR	ID	BR	TH	MX	MY	RU	GB	IN
US	0.73786	0.00035	0.00610	0.00156	0.00095	0.00078	0.00049	0.00024	0.01234	0.00125
TR	0.00035	0.00376	0.00003	0.00000	0.00000	0.00000	0.00000	0.00001	0.00010	0.00000
ID	0.00610	0.00003	0.12443	0.00003	0.00016	0.00001	0.00062	0.00002	0.00069	0.00008
BR	0.00156	0.00000	0.00003	0.02145	0.00000	0.00004	0.00001	0.00003	0.00021	0.00002
TH	0.00095	0.00000	0.00016	0.00000	0.01421	0.00000	0.00006	0.00000	0.00020	0.00002
MX	0.00078	0.00000	0.00001	0.00004	0.00000	0.00396	0.00000	0.00000	0.00005	0.00001
MY	0.00049	0.00000	0.00062	0.00001	0.00006	0.00000	0.00768	0.00001	0.00016	0.00004
RU	0.00024	0.00001	0.00002	0.00003	0.00000	0.00000	0.00001	0.00283	0.00006	0.00001
GB	0.01234	0.00010	0.00069	0.00021	0.00020	0.00005	0.00016	0.00006	0.02568	0.00020
IN	0.00125	0.00000	0.00008	0.00002	0.00002	0.00001	0.00004	0.00001	0.00020	0.00424



(a) (b) (c)

FIGURE 4. Analyzing the swarm social graph. (a) Size of connected components. (b) Path length (LCC). (c) Robustness.

3) PAGERANK

PageRank is a metric that quantifies the importance of different nodes in the network [40]. It has been applied by the Google search engine to rank the websites. For any node of the network, its PageRank value is between 0 and 1. A larger PageRank value indicates that the corresponding node is more important. PageRank has been widely used in quantifying the user influence [27], [45], [46], [49] in social networks. Based on the Swarm social graph, we are able to compute the PageRank of all nodes. We define the users within the top 0.1% PageRank values as *influentials*, and use a set P to present them. We aim to find some unique characteristics of the influentials.

Regarding the graph metrics, the average degree of nodes in P is 655.30, which is much larger than that of the whole network. Meanwhile, the average clustering coefficient is 0.041, which is instead much smaller than that of the whole network. We believe this is because the influentials are globally well connected and their friends are not densely connected with each other.

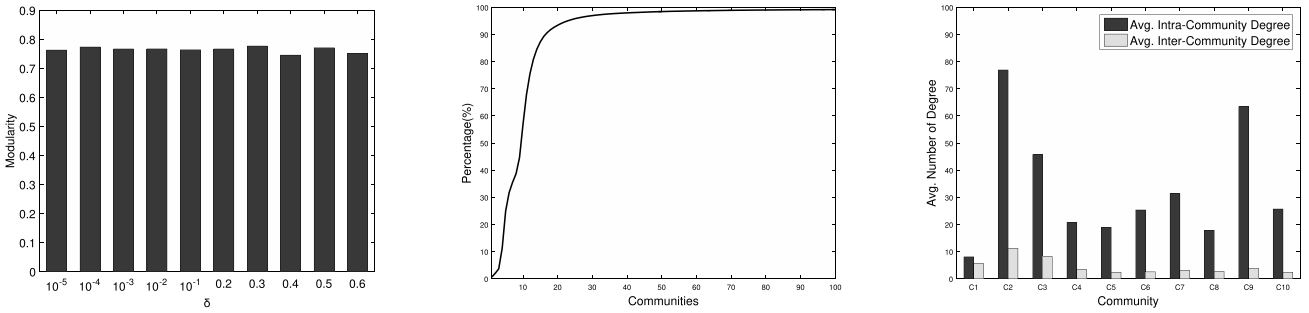
Also, we can see the influentials has a different composition of gender and country. In P, 68.55% of users are male, 26.94% are female, and the rest do not provide any gender information. Therefore, there are more male users in P. Regarding the country composition, we can see that the top three countries are USA (34.42% of all users), Indonesia (7.90%), and Russia (4.80%). Regarding the cross-site linking option, 87.32% of the influentials have enabled this option, which is much larger than that of all Swarm users (56.76%).

4) CONNECTED COMPONENTS

A big social graph might have a number of connected components. A connected component is an undirected subgraph. In a connected component, any two nodes are connected to each other by paths, and any of these nodes is not connected to any additional node in the supergraph. There are 35,690,080 nodes (57.01% of all nodes) in the LCC, or, in the giant component. Also, 42.50% of Swarm nodes are singletons [25], i.e., nodes with zero degree. Among the nodes with a non-zero degree, only 305,064 of them do not belong to the LCC. The exceptional cases cover only 0.49% of all Swarm nodes. As in [25], we call these nodes "middle region". As shown in Fig. 4(a), there is a single giant connected component in the Swarm network. The sizes of the second, third, fourth, and fifth largest connected components are 1614, 407, 301, and 301, respectively.

Fig. 4(b) shows the probability density function of the shortest-path distances of node pairs in the LCC. We can see the average distance is only 5.12. The 90th percentile value of the distances is 6, which is also known as the effective diameter [28] of the LCC. This means for most of the node pairs in the LCC, they can reach each other within 6 hops.

As in [32], we further examine whether the "core" of the Swarm network is densely connected. We remove from the entire social graph the nodes with the highest degrees, and analyze the remaining nodes and edges. Starting from 0.01%, we remove up to 30% of the nodes with the highest degrees. The distribution of the connected components in the remained network is illustrated in Fig. 4(c). Note that we classify all connected components into three groups, i.e., the LCC,



(a) (b) (c)

FIGURE 5. Community structure of the swarm network. (a) Modularity v.s. δ. (b) Coverage of top 100 communities. (c) Communities edges.

the singletons, and the middle region. We can see that even when 20% of highest degree nodes have been removed, we still have a giant LCC, covering over 30% of the remaining nodes. Obviously, the graph is still well connected even when we remove 20% of the nodes with the highest degrees. Once we remove 30% of these nodes, the LCC is split into a large number of connected components with few nodes in each component. We can see a much smaller LCC, and the percentage of the middle region nodes has become much larger.

5) COMMUNITY DETECTION

"Communities" widely exist in different types of online social networks [26], [54]. A community is a group of densely connected nodes. The inter-community connections are relatively sparse. For the Swarm network, we study how users form communities in such a large network. Here we exclude singletons, as each of them will form a single-node community. We focus on the rest 35,995,144 nodes and all the edges between them.

To detect the communities in the Swarm network, we use the Louvain algorithm [2], which has been applied for different types of networks [16], [18], [54]. This algorithm can scale to a network with tens of millions of nodes. To evaluate the accuracy of the community detection, a metric known as *modularity* has been widely used. The value of modularity is between -1 and 1. It measures the density of intra-communities links as compared to inter-community links. Precisely, if there are k communities, modularity Q can be calculated as

$$Q = \sum_{i=1}^{k} (e_{ii} - a_i^2) \tag{3}$$

In Eq. 3, e_{ij} refers to the fraction of edges with one end nodes in community i and the other in community j, while a_i represents the fraction of ends of edges that are attached to nodes in community i, i.e., $a_i = \sum_j e_{ij}$. A larger modularity indicates the network can be clustered into communities in a better way. According to [26], modularity ≥ 0.3 means the corresponding network has a viable community structure. As discussed in [54], the δ parameter is a critical tuning parameter for the Louvain algorithm. According to Fig. 5(a), we can see that different δ values will lead to a similar modularity. Moreover, among all δ values we have chosen, the corresponding modularity values are always much larger than 0.3. Therefore, the Swarm users can be grouped into communities well.

TABLE 5. Percentage of users in top 3 countries per community.

Community (size)	Countries (% of Users)		
$C1$ $(6,538,135)$	US(75.74%)	GB(4.17%)	NL(3.00%)
$C2$ $(4,792,055)$	TR(93.39%)	US(2.99%)	DE(0.48%)
$C3$ $(3,301,595)$	BR(79.53%)	US(6.72%)	CO(4.89%)
$C4$ $(2,781,426)$	ID(78.85%)	US(11.57%)	NY(3.31%)
$C5$ $(1,879,103)$	RU(52.03%)	US(17.16%)	UA(14.26%)
$C6$ $(1,278,430)$	MX(70.22%)	US(10.58%)	AR(9.39%)
$C7$ $(1,157,494)$	MY(81.17%)	US(8.42%)	NY(2.19%)
$C8$ $(994,740)$	TH(80.96%)	US(11.81%)	NY(1.42%)
$C9$ $(726,985)$	TR(93.05%)	US(2.93%)	DE(0.42%)
$C10$ $(618,689)$	PH(68.23%)	US(13.68%)	AE(2.14%)

Given the similarity among modularity values, we set δ as 0.1 for our further study. We can see that the investigated nodes can be clustered into 137,470 communities. Although the number of communities is large, only a few of them have many nodes. We plot the cumulative node percentage of the largest 100 communities in Fig. 5(b). We can see the top 20 communities have covered 86.3% of all investigated nodes, and top 100 communities have covered 99.6% of all investigated nodes. Therefore, most of the nodes belong to a small number of communities.

For an edge connecting two nodes within the same community, we denote it as an "intra-community" edge. Differently, for an edge connecting two nodes from different communities, we call it an "inter-community" edge. In Fig. 5(c), for each of the top 10 communities, we show the average intra-community degree and inter-community degree. We can see that for all these 10 communities, the value of the average intra-community degree is significantly larger than that of the average inter-community degree.

We further dive into the largest communities, and study the relationship between these communities and the users' home countries. According to Table 5, each community is dominant by one or a small number of countries. In other words, users

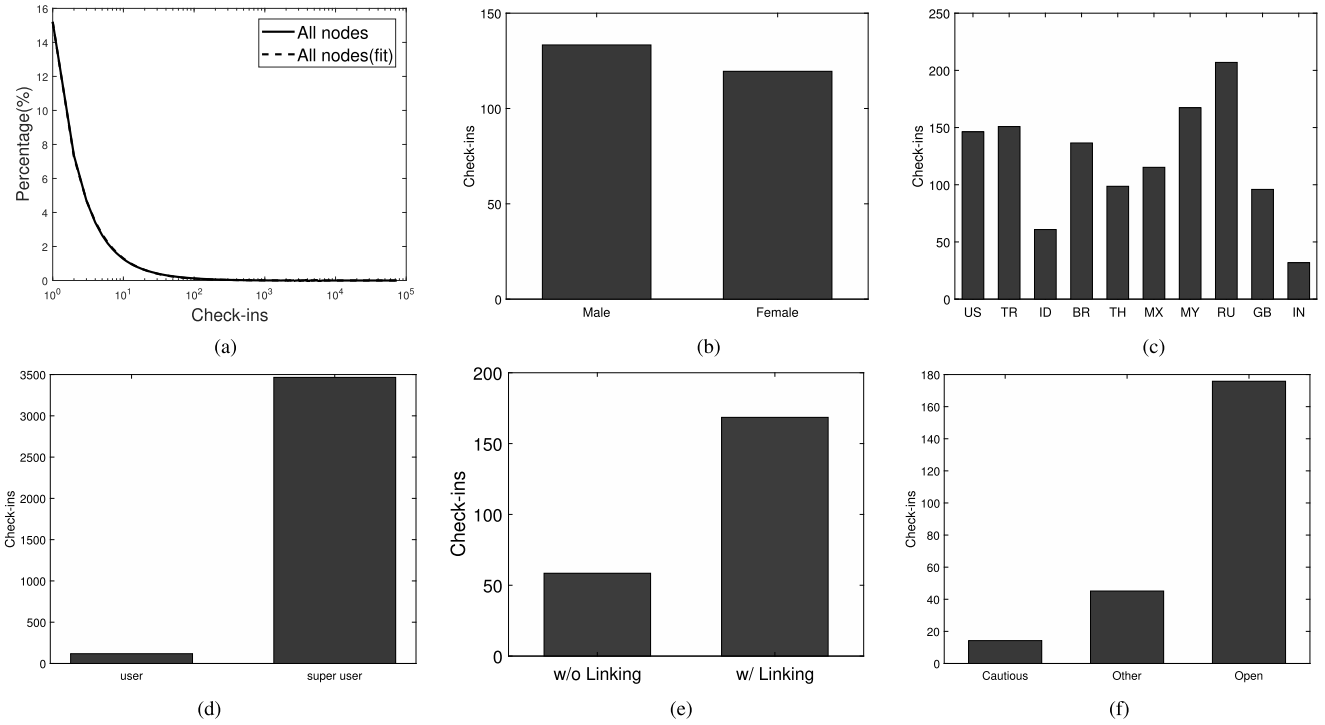


FIGURE 6. Check-ins of swarm users. (a) Probability distribution function. (b) Gender. (c) Country. (d) User type. (e) Cross-site linking. (f) Privacy.

have a higher chance to connect with other users coming from the same country. Moreover, the dominant countries in the top four communities are United States, Turkey, Brazil and Indonesia, respectively. As we have shown in Section III-A, these four countries have the largest Swarm population.

6) SUMMARY

In short, the Swarm social graph has an average node degree of 24.84, and the degrees reveal a clear assortative mixing pattern. The nodes in this graph are loosely connected with each other. More than 57% of all the nodes belong to the LCC, and the users can be divided well into communities.

D. GO BEYOND THE GRAPH: CHECK-INS OF SWARM USERS

In this subsection, we focus on the check-in function, which is the core function of the Swarm app, and the principal form of user-generated contents (UGCs). A user can share her latest check-ins with her Swarm friends. Meanwhile, users can compete for the ''mayor'' for a certain place, by conducting more check-ins. This motivates the users to post more. In this subsection, we investigate the check-in activities by referring to different groups of Swarm users, and study the relationship between a user's activeness in check-ins and the user's profile and social connections.

Fig. 6(a) shows the probability density function of the number of check-ins of all Swarm users. About 44.88% of users have conducted check-ins, while a Swarm user has undertaken 120.94 check-ins on average. Top 1% of Swarm users have performed 34.73% of all check-ins, and top 5% of Swarm users have conducted 70.26% of all check-ins. We also find that the number of check-ins follows a power-law distribution ($C = 15.21, \alpha = 1.068$). The corresponding R^2 value is 0.9997.

We classify users into groups. According to Fig. 6(b), male users have published on average 133.34 check-ins, while female users have posted 119.51 check-ins. Therefore, male users are more active in conducting check-ins. Differently, in microblogging networks like Twitter, female users publish more tweets than male users [8]. Among different countries in Fig. 6(c), users from Russia and Malaysia are more active in performing check-ins. On average, a user in Russia has published 206.96 check-ins, while in India, this number is only 60.84. According to Fig. 6(d), there is a significant difference between superusers and ordinary users, in terms of the number of published check-ins. As shown in Fig. 6(e), enabling the cross-site linking function indicates a larger number of check-ins in general. This is consistent with our earlier findings in [8]. Finally, as shown in Fig. 6(f), open users tend to publish more, while cautious users would prefer to publish less.

Based on the number of published check-ins, we are able to classify all Swarm users into two groups, i.e., a group of ''active users'' and another group of ''less active users''. By examining the data set, we find that each of the top 10% of users has conducted more than 220 check-ins, and each of the top 20% of users has posted more than 55 check-ins. Accordingly, if a user has conducted more than 55 check-ins, we will consider her as an ''active user''. Otherwise, this user will be selected as a ''less active user''. We further investigate

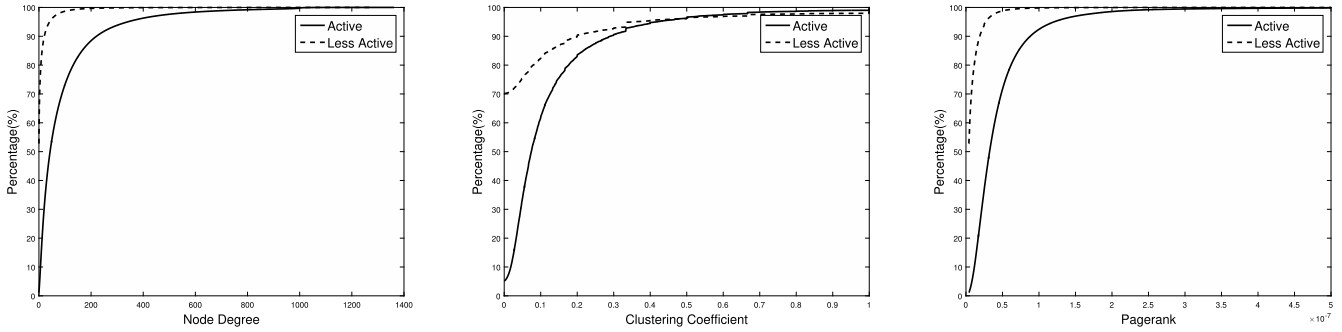


FIGURE 7. Comparison between Active and Less Active users based on graph metrics. (a) Degree. (b) Clustering coefficient. (c) PageRank.

TABLE 6. Comparison between active and less active users based on optional profile fields.

	Gender	Lastname	Biography	Profile Photo	Facebook	Twitter	Location
Active	97%	92%	12%	96%	72%	44%	93%
Less active	91%	94%	2%	61%	48%	8%	89%

that to what extent a user's social connections and optional profile fields will affect the classification of users. In other words, we investigate the feasibility of predicting whether a user is active or not based on her social connections and optional profile fields.

In practice, we apply supervised machine learning techniques to train a classifier. We select a number of key features to distinguish users from different groups. There are two categories of these features as follows.

- Graph metrics (4 features): We use three key graph metrics as features, including the degree, clustering coefficient, and the PageRank values of a user. Also, we use the community number of a user as one of the features. For a singleton, the corresponding community number will be set as -1.
- Profile fields (7 features): We pick several optional fields in a users profile, and see whether each field is enabled or not. These fields include gender, location, profile photo, Facebook account, Twitter account, biography, and last name. If a field is enabled, we set the feature value as 1. Otherwise, we set the feature value as 0.

We randomly pick 2000 active users and 2000 less active users to form a training and validation data set. Using some features related to the social graph, we can see significant differences between active users and inactive users. In Fig. 7, we compare active users and less active users in terms of degree, clustering coefficient and PageRank. Similarly, we compare active users and less active users by checking whether a certain optional profile field in Table 6 is enabled. We find significant differences between these two user groups in profile fields of biography, profile photo, Facebook account and Twitter account.

We use a number of classic machine learning algorithms, including XGBoost [6], support vector machine (SVM) [19], C4.5 Decision Tree [42], Random Forest [3], and Naive Bayers [24]. XGBoost is an emerging scalable tree boosting system. It has been widely used in different machine learning contests. Besides XGBoost, the other algorithms are evaluated using Weka [17]. For SVM, we apply both SVM with radial basis function kernel (SVMr) and SVM with polynomial kernel (SVMp).

To evaluate the prediction performance of the classifiers, we apply the following three metrics, i.e., precision, recall, and F1-score. Precision means the fraction of predicted active users who are really active. Recall means the fraction of active users who are accurately detected. As in Eq. 4, F1-score is defined as the harmonic mean of precision and recall.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

Once a set of parameters are determined, we can use 10-fold cross-validation[4] to calcualte the precision, recall and F1-score. For each algorithm, we carefully tune the parameters, and record a set of "best" parameters which achieves the maximum F1-score. Please refer to Tables 7 and 8 for the parameters.

After the models have been trained, we randomly select 1000 active users and 1000 less active users to form a test data set. We use the trained models to predict each user's category.

According to our results shown in Table 7, we have found that XGBoost performs the best with an F1-score of 0.883. Therefore, we can conclude that the selected features could accurately distinguish active users from less active users.

To measure different features' discriminative power, we show the top 10 features ranked by χ^2 (Chi Square) statistic [52]. As shown in Table 9, we can see that the most discriminative features are PageRank, degree, clustering coefficient, and the community information, which are all

[4]In 10-fold cross-validation, the training and validation data set is randomly divided into 10 subsets with equal size. Of the 10 subsets, a single subset is retained as the validation data for evaluating the model, and the remaining 9 subsets are used for training. The cross-validation process is repeated 10 times, with each of the 10 subsets used once as the validation data.

TABLE 7. Prediction of active users.

Algorithm	Parameter	Precision	Recall	F1-score
XGBoost	Refer to Table VIII	0.924	0.846	0.883
RandomForest	100 trees, 7 features/tree, maximum depth=10	0.830	0.915	0.871
C4.5 (J48)	Confidence factor C =0.25, Instance/leaf M =2	0.835	0.900	0.866
NaiveBayes	default	0.792	0.923	0.853
SVMr	Kernel γ =0.0625, Cost parameter C =256	0.859	0.857	0.858
SVMp	Kernel degree d =3, Cost parameter C =32	0.848	0.848	0.848

TABLE 8. Parameters set for XGBoost.

Parameter	*Value*
learning_rate	0.1
min_child_weight	9
max_depth	4
gamma	3.1
subsample	0.35
colsample_bytree	0.95
lambda	1
alpha	0
booster	gbtree
objective	binary:logistic

TABLE 9. χ^2 statistic.

Rank	Feature	χ^2
1	PageRank	9140.08
2	Degree	8481.52
3	Clustering coefficient	7558.13
4	Community number	5340.25
5	Profile Photo	2845.97
6	Twitter	2630.99
7	Facebook	952.22
8	Biography	509.46
9	Gender	261.67
10	Location	70.63

determined by the social graph. Therefore, the social connections and optional profile fields of a Swarm user could provide a viable hint for her check-in activities. Intuitively, we believe that this is because the network will spread the check-in information to a user's friends via the social network. Therefore, the social network encourages Swarm users to conduct more check-ins.

IV. RELATED WORK

Social network analytics has become a widely used tool to understand the connections among OSN users [1], [21], [32], [50], [54]. Nowadays the sizes of the OSNs are growing rapidly. This makes the analysis of the entire social graph of a mainstream OSN a challenging problem. Ugander *et al.* [47] have studied the social network of active users of Facebook in May 2011, covering 721 million users. Myers *et al.* [34] have analyzed the social graph of Twitter in 2012 with 175 million active users and about 20 billion edges. Unfortunately, these data sets are provided by the OSN service providers directly. In most cases, researchers from the academia or third-party application providers are not able to access such data sets, and have to crawl the publicly-accessible data from the OSN sites. Gabielkov *et al.* [12] aimed to crawl the entire Twitter population, and study the macroscopic anatomy of the Twitter social graph. However, more than 5% of Twitter users are "protected", i.e., their connections to other users are unavailable to the public. As a result, the entire Twitter graph cannot be collected by crawling the public data. Gong *et al.* [14] have crawled the Largest Weakly Connected Component (LWCC) of Google+. The crawled data set covers about 70% of all users. Similarly, Gonzalez *et al.* [15] have crawled five snapshots of LWCC of Google+ using BFS. Still, the entire Google+ social graph has not been obtained.

Given the difficulty in obtaining an entire network with millions of users, people might crawl a subset of such networks, and investigate the crawled subset such as [21], [32], and [50]. There are several proposals on how to sample a representative subset, such as BFS, Metropolis-Hastings Random Walk [13] and Frontier Sampling [43]. However, according to our earlier study in [48], none of these algorithms can preserve all key properties of the original graph. As we have shown in Section III-B, getting the entire graph is necessary to conduct a comprehensive and accurate analysis of key properties of the graph.

Due to the rapid development of mobile computing technologies, a number of social networking services, such as WeChat, Swarm, Momo and WhatsApp, are only available in mobile platforms. However, there are very few work on conducting data-driven analysis of these networks. Huang *et al.* [20] have collected WeChat traces from a commercial 3G network in China, including about 150K WeChat users, and further investigated the user and service/task activities. Chen *et al.* [7] have crawled over 8 million user profiles and around 150 million location updates from Momo, and studied the spatial-temporal usage patterns of Momo users. Fiadino *et al.* [11] have analyzed an entire week of WhatsApp traffic traces collected at the core of an European nation-wide cellular network, and presented a large-scale traffic characterization of WhatsApp. Noulas *et al.* [39] have crawled a data set of about 700K Foursquare users, and studied spatial-temporal patterns of these users. Qiu *et al.* [41] have explored social messaging groups of WeChat, focusing on the lifecycle, the change in group structures over time and the membership cascade process. Their study is based on an anonymized data set provided by WeChat. Differently, our work has demonstrated how to efficiently crawl and analyze an entire mobile social network with more than 60 million users around the world, and how to analyze this network from different aspects.

V. CONCLUSION AND FUTURE WORK

In this paper, we have conducted a data-driven analysis for the entire Swarm network. Our study covers 62.6 million Swarm users, and we investigate the social connections among them. By analyzing such a huge social graph, we use several classic graph metrics to characterize how people in Swarm connect

with each other. Furthermore, we study the check-ins, the primary form of UGCs in Swarm. We use some graph-based and profile-based features to accurately predict the check-in activities. For future work, we aim to study the following issues.

First, as we have obtained the entire social graph of a mobile social network, we will use this graph as the "ground truth" to evaluate different applications related to big social graphs. Potential applications include large-scale graph processing systems [31], and embedding a huge graph into geometric spaces [53]. Also, we could study the malicious account detection problem, by using social graph-based algorithms such as SybilRank [4].

Second, in this paper, we only study the number of check-ins, as detailed check-in history of a user is not available to the public. To explore further into the spatial-temporal properties of Swarm users, we need the detailed check-in history of some users. As in [30], we will recruit some volunteers who could give us access to their check-in data, and we plan to further investigate the relationship between social connections and user mobility.

Last but not least, we aim to understand the dynamic natures of the Swarm network. We plan to do the data crawling periodically to obtain a series of snapshots of the network. We will further study how the network evolves. Meanwhile, we are interested in the link prediction problem of the Swarm network. We aim to develop some algorithms to predict potential social connections in an accurate way.

REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *Proc. WWW*, 2007, pp. 835–844.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Statist. Mech., Theory Experim.*, vol. 2008, no. 10, p. P10008, 2008. [Online]. Available: http://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008/meta

[3] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[4] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *Proc. USENIX/ACM NSDI*, 2012, p. 15.

[5] S. Caton, C. Haas, K. Chard, K. Bubendorfer, and O. F. Rana, "A social compute cloud: Allocating and sharing infrastructure resources via social networks," *IEEE Trans. Services Comput.*, vol. 7, no. 3, pp. 359–372, Jul./Sep. 2014.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM KDD*, 2016, pp. 785–794.

[7] T. Chen, M. A. Kaafar, and R. Boreli, "The where and when of finding new friends: Analysis of a location-based social discovery network," in *Proc. AAAI ICWSM*, 2013, pp. 61–70.

[8] Y. Chen, C. Zhuang, Q. Cao, and P. Hui, "Understanding cross-site linking in online social networks," in *Proc. 8th Workshop Social Netw. Mining Anal. (SNAKDD)*, 2014, Art. no. 6.

[9] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.

[10] C. Ding, Y. Chen, and X. Fu, "Crowd crawling: Towards collaborative data collection for large-scale online social networks," in *Proc. ACM COSN*, 2013, pp. 183–188.

[11] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting WhatsApp in cellular networks: Servers, flows, and quality of experience," in *Proc. 7th Int. Workshop Traffic Monitoring Anal.*, 2015, pp. 49–63.

[12] M. Gabielkov, A. Rao, and A. Legout, "Studying social networks at scale: Macroscopic anatomy of the twitter social graph," in *Proc. ACM SIGMETRICS*, 2014, pp. 277–288.

[13] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A case study of unbiased sampling of OSNs," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

[14] N. Z. Gong *et al.*, "Evolution of social-attribute networks: Measurements, modeling, and implications using Google+," in *Proc. ACM IMC*, 2012, pp. 131–144.

[15] R. Gonzalez, R. Cuevas, R. Motamedi, R. Rejaie, and A. Cuevas, "Google+ or Google−?: Dissecting the evolution of the new OSN in its first year," in *Proc. WWW*, 2013, pp. 483–494.

[16] D. Greene, D. Doyle, and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *Proc. ASONAM*, 2010, pp. 176–183.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.

[18] J. Haynes and I. Perisic, "Mapping search relevance to social networks," in *Proc. 3rd Workshop Social Netw. Mining Anal.*, 2009, Art. no. 2.

[19] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.

[20] Q. Huang, P. P. C. Lee, C. He, J. Qian, and C. He, "Fine-grained dissection of WeChat in cellular networks," in *Proc. IEEE 23rd Int. Symp. Quality Service (IWQoS)*, Jun. 2015, pp. 309–318.

[21] J. Jiang *et al.*, "Understanding latent interactions in online social networks," in *Proc. ACM IMC*, 2010, pp. 369–382.

[22] L. Jiao, J. Li, T. Xu, W. Du, and X. Fu, "Optimizing cost for online social networks on geo-distributed clouds," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 99–112, Feb. 2016.

[23] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos, "Understanding user behavior in online social networks: A survey," *IEEE Commun. Mag.*, vol. 51, no. 9, pp. 144–150, Sep. 2013.

[24] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proc. UAI*, 1995, pp. 338–345.

[25] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Proc. ACM KDD*, 2006, pp. 611–617.

[26] H. Kwak, Y. Choi, Y.-H. Eom, H. Jeong, and S. Moon, "Mining communities in networks: A solution for consistency and its evaluation," in *Proc. ACM IMC*, 2009, pp. 301–314.

[27] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proc. WWW*, 2010, pp. 591–600.

[28] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *Proc. ACM KDD*, 2005, pp. 177–187.

[29] J. Leskovec and R. Sosič, "SNAP: A general-purpose network analysis and graph-mining library," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 1, 2016, Art. no. 1.

[30] S. Lin, R. Xie, Q. Xie, H. Zhao, and Y. Chen, "Understanding user activity patterns of the swarm app: A data-driven study," in *Proc. ACM UbiComp/ISWC*, 2017, pp. 125–128.

[31] G. Malewicz *et al.*, "Pregel: A system for large-scale graph processing," in *Proc. ACM SIGMOD*, 2010, pp. 135–146.

[32] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. ACM IMC*, 2007, pp. 29–42.

[33] A. Mohaisen, H. Tran, A. Chandra, and Y. Kim, "Trustworthy distributed computing on social networks," *IEEE Trans. Serv. Comput.*, vol. 7, no. 3, pp. 333–345, Jul./Sep. 2014.

[34] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network?: The structure of the twitter follow graph," in *Proc. WWW Companion*, 2014, pp. 493–498.

[35] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett.*, vol. 89, no. 20, p. 208701, 2002.

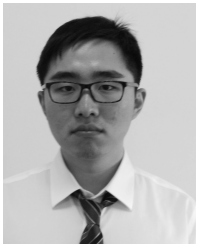
[36] M. E. J. Newman, "Mixing patterns in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 67, no. 2, p. 026126, 2003.

[37] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 68, no. 3, p. 036122, 2003.

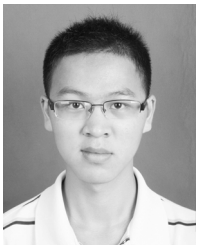
[38] A. Noulas, R. Lambiotte, B. Shaw, and C. Mascolo, "Topological properties and temporal dynamics of place networks in urban environments," in *Proc. WWW*, 2015, pp. 431–441.

[39] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare," in *Proc. AAAI ICWSM*, 2011, pp. 570–573.
[40] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the Web," Stanford Univ., Stanford, CA, USA, Tech. Rep. SIDL-WP-1999-0120, 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/cgi/export/422/BibTeX/ilprints-eprint-422.bib
[41] J. Qiu *et al.*, "The lifecycle and cascade of WeChat social messaging groups," in *Proc. WWW*, 2016, pp. 311–320.
[42] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, 1993.
[43] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multi-dimensional random walks," in *Proc. ACM IMC*, 2010, pp. 390–403.
[44] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks," in *Proc. AAAI ICWSM*, 2011, pp. 329–336.
[45] X. Song, Y. Chi, K. Hino, and B. Tseng, "Identifying opinion leaders in the blogosphere," in *Proc. ACM CIKM*, 2007, pp. 971–974.
[46] J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogenous networks," in *Proc. ACM WSDM*, 2012, pp. 743–752.
[47] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the Facebook social graph," *CoRR*, Nov. 2011. [Online]. Available: https://arxiv.org/abs/1111.4503
[48] T. Wang *et al.*, "Understanding graph sampling algorithms for social network analysis," in *Proc. IEEE ICDCS Workshops*, Jun. 2011, pp. 123–128.
[49] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: Finding topic-sensitive influential twitterers," in *Proc. ACM WSDM*, 2010, pp. 261–270.
[50] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proc. ACM EuroSys*, 2009, pp. 205–218.
[51] T. Xu, Y. Chen, L. Jiao, B. Y. Zhao, P. Hui, and X. Fu, "Scaling microblogging services with divergent traffic demands," in *Proc. 12th Int. Middleware Conf.*, 2011, pp. 20–40.
[52] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. ICML*, 1997, pp. 412–420.
[53] X. Zhao, A. Chang, A. D. Sarma, H. Zheng, and B. Y. Zhao, "On the embeddability of random walk distances," *VLDB Endow.*, vol. 6, no. 14, pp. 1690–1701, 2013.
[54] X. Zhao *et al.*, "Multi-scale dynamics in a massive online social network," in *Proc. ACM IMC*, 2012, pp. 171–184.
[55] C. Zhong, M. Salehi, S. Shah, M. Cobzarenco, N. Sastry, and M. Cha, "Social bootstrapping: How pinterest and last.fm social communities benefit by borrowing links from Facebook," in *Proc. WWW*, 2014, pp. 305–314.
[56] J. Zhou and J. Fan, "JPR: Exploring joint partitioning and replication for traffic minimization in online social networks," in *Proc. IEEE ICDCS*, Jun. 2017, pp. 1147–1156.

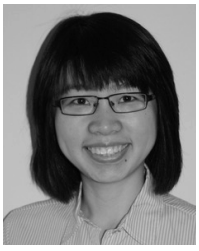
YANG CHEN (M'07–SM'15) received the B.S. and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, in 2004 and 2009, respectively. He visited Stanford University in 2007 and Microsoft Research Asia from 2006 to 2008 as a Visiting Student. From 2009 to 2011, he was a Research Associate and the Deputy Head of the Computer Networks Group, Institute of Computer Science, University of Goettingen, Germany. From 2011 to 2014, he was a Post-Doctoral Associate with the Department of Computer Science, Duke University, Durham, NC, USA, where he served as a Senior Personnel with the NSF MobilityFirst Project. He is currently a Pre-Tenure Associate Professor with the School of Computer Science, Fudan University, where he is also the Leader of the Mobile Systems and Networking Group. He published more than 50 referred papers in international journals and conferences, including IEEE TPDS, IEEE TSC, IEEE TNSM, the *IEEE Communications Magazine*, Middleware, INFOCOM, ICDE, COSN, CIKM, and IWQoS. His research interests include online social networks, Internet architecture, and mobile computing. He served as an OC/TPC Member for many international conferences, including SOSP, WWW, IJCAI, AAAI, IWQoS, ICCCN, GLOBECOM, and ICC. He is serving as an Associate Editor of the IEEE Access and an Editorial Board Member of the *Transactions on Emerging Telecommunications Technologies*.



JIYAO HU received the B.S. degree from the School of Computer Science, Fudan University, in 2017. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Duke University. He was a Student Research Assistant with the Mobile Systems and Networking Group from 2014 to 2017. He visited Aalto University as a Research Intern in 2017. His research interests include online social networks and data mining.



Hao Zhao is currently pursuing the degree from the School of Computer Science, Fudan University. He has been a Student Research Assistant with the Mobile Systems and Networking Group since 2015. He visited Tsinghua University in 2016 and the University of Goettingen in 2017 as a Research Intern. His research interests include massive data analytics and machine learning.



Yu Xiao received the bachelor's and master's degrees in computer science and technology from the Beijing University of Posts and Telecommunications, China, and the Ph.D. degree (Hons.) in computer science from Aalto University in 2012. She is currently an Assistant Professor with the Department of Communications and Networking, Aalto University, where she also the Leader of the Mobile Cloud Computing Group. Her research interests include edge computing, mobile crowdsensing, and energy-efficient wireless networking. Her work has received three best paper awards from IEEE/ACM conferences. She was a recipient of the three-year post-doctoral grant from the Academy of Finland.

Pan Hui (M'04–SM'14–F'18) received the M.Phil. and B.Eng. from the Department of Electrical and Electronic Engineering, The University of Hong Kong, and the Ph.D. degree from the Computer Laboratory, University of Cambridge. He has been the Nokia Chair in data science and a Full Professor in computer science with the University of Helsinki since 2017. He has been a Faculty Member with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, since 2013, and an Adjunct Professor of social computing and networking with Aalto University, Finland, since 2012. He was a Senior Research Scientist and a Distinguished Scientist for Telekom Innovation Laboratories, Germany, from 2008 to 2015. During his Ph.D. period, he was with Intel Research Cambridge and Thomson Research Paris. He has published over 200 research papers with over 12 500 citations and has around 30 granted/filed European patents. He has founded and chaired several IEEE/ACM conferences/workshops. He has been serving on the organizing and Technical Program Committee of numerous international conferences, including ACM SIGCOMM, IEEE Infocom, ICNP, SECON, MASS, Globecom, WCNC, ITC, IJCAI, ICWSM, and WWW. He is an Associate Editor for the IEEE Transactions on Mobile Computing and the IEEE Transactions on Cloud Computing, a Guest Editor for the *IEEE Communication Magazine*, and an ACM Distinguished Scientist.